Fast Approximate Model for the 3D Matter Power Spectrum

Arrykrishna Mootoovaloo[*] **Andrew H. Jaffe , Alan F. Heavens and Florent Leclercq**
Imperial Centre for Inference and Cosmology (ICIC), Department of Physics
Imperial College London
Prince Consort Road, SW7 2AZ
{a.mootoovaloo17, a.jaffe, a.heavens, f.leclercq}@imperial.ac.uk

Abstract

Many Bayesian inference problems in cosmology involve complex models. Despite the fact that these models have been meticulously designed, they can lead to intractable likelihood and each forward simulation itself can be computationally expensive, thus making the inverse problem of learning the model parameters a challenging task. In this paper, we develop an approximate model for the 3D matter power spectrum, $P_\delta(k, z)$, which is a central quantity in a weak lensing analysis. An important output of this approximate model, often referred to as surrogate model or emulator, are the first and second derivatives with respect to the input cosmological parameters. Without the emulator, the calculation of the derivatives requires multiple calls of the simulator, that is, the accurate Boltzmann solver, CLASS (Lesgourgues, 2011). We illustrate the application of the emulator in the calculation of different weak lensing and intrinsic alignment power spectra and we also demonstrate its performance on a toy simulated weak lensing dataset.

1 Introduction

The 3D matter power spectrum, $P_\delta(k, z)$ is central to most cosmological data analysis, such as cosmic shear, galaxy clustering and others. These analyses require the computation of relevant power spectra and the latter can be computed numerically in a fast way only if the calculation of $P_\delta(k, z)$ itself is fast enough. Machine Learning (ML) techniques have been exploited to accelerate cosmological data analysis. For example, recent techniques such as density estimation uses the Expectation-Maximisation (EM) and neural networks (NN) to directly learn the posterior distribution of cosmological and nuisance parameters from a set of compressed data vectors, evaluated at different points in the parameter space (Alsing et al., 2018, 2019; Alsing & Wandelt, 2019). These techniques can further be explored and applied to other cosmological analysis, for example, a weak lensing analysis.

The surrogate model developed in this work can easily pave its way in current and future weak lensing analysis. A weak lensing analysis requires the calculation of $\frac{1}{2}n(n+1)$ weak lensing and intrinsic alignment power spectra and in recent analyses, 5 tomographic redshift distributions are used and this results in the calculation of 15 power spectra. In future surveys, it is expected that the number of tomographic redshift distributions will increase to 10 and this will be very expensive if we use standard solvers such as CAMB (Lewis & Challinor, 2011) or CLASS. However, the different power

[*] https://harry45.github.io/

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

spectra can easily be calculated once $P_{\delta}(k,\chi)$ is computed since they all involve integration of the form:

$$C_{\ell}=\int_{0}^{\chi_{\mathrm{H}}} g(\chi)\,P_{\delta}(k,\chi)\,\mathrm{d}\chi. \tag{1}$$

χ is the comoving radial distance and $g(\chi)$ is a function of the redshift distribution, $n(\chi)$. $P_{\delta}(k,\chi)$ becomes more expensive if we choose to use large N-body simulations, where each forward simulation can take minutes or hours.

Related Work: Emulation has been applied in different cosmological data analyses. For example, Fendt & Wandelt (2007) developed the Parameters for the Impatient Cosmologist (PICO), which is based on polynomial regression, to interpolate Cosmic Microwave Background (CMB) power spectra at test points in parameter space. A similar application to CMB was performed by Auld et al. (2007) using neural networks. Gaussian Process (GP) was used in the Coyote Universe collaboration (Habib et al., 2007; Heitmann et al., 2009, 2010, 2014; Lawrence et al., 2010) for emulating the matter power spectrum for N-body simulations. In the same spirit, other emulators were designed based on neural networks for the matter power spectrum, 21cm power spectrum in the context of epoch of reionisation and others (Agarwal et al., 2012, 2014; Aricò et al., 2021; Ho et al., 2021).

In this work, we re-write the 3D matter power spectrum as

$$P_{\delta}(k,z)=D(z)[1+q(k,z)]P_{\text{lin}}(k,z_{0}) \tag{2}$$

where $D(z)$ is the linear growth factor (assumed scale-independent), and $P_{\text{lin}}(k,z_0)$ is a scale-independent reference linear matter power spectrum at fixed redshift z_0. The quantity $q(k,z)$ not only encapsulates the non-linear contributions, but also any scale-dependence in the linear growth factor, for instance due to massive neutrinos or modified gravity. We first approximate the 3 different quantities at each redshift, z and wavenumber, k, with a set of polynomial functions and we model the residuals using a kernel function. Once the model is trained and stored, we can also compute analytical first and second derivatives of $P_{\delta}(k,z)$. Moreover, following Equation 1, we can now use the surrogate model to compute any power spectra and in the weak lensing context, we compute three different power spectra. These are then used to in a Bayesian inverse problem to infer cosmological parameters using a simulated dataset.

2 The Approximate Model

Following a recent weak lensing analysis by Köhlinger et al. (2017), we follow a similar range for all cosmological parameters

$$\boldsymbol{\theta}=[\Omega_{\text{cdm}}h^{2},\,\Omega_{\text{b}}h^{2},\,\ln(10^{10}A_{s}),\,n_{s},\,h]$$

and the redshift range is $z\in[0.0,\,5.0]$ and the wavenumber range, $k\in[10^{-4},\,50.0]$. The input training points are generated using Latin Hypercube Sampling (LHS) and these are scaled to the appropriate range of the input cosmological parameters. $P_{\delta}(k,z)$ is then evaluated at 20 redshifts in the linear scale and at 40 wavenumbers in the logarithmic scale corresponding to the pre-defined ranges. The three different components, namely the growth factor, $D(z)$, the $q(k,z)$ function and the linear matter power spectrum $P_{\text{lin}}(k,z_0)$ are then computed at N design points, $\boldsymbol{\theta}$, such that we have a training set, $\{\boldsymbol{\theta},\boldsymbol{y}_i\}$. This results in a total of 860 outputs from the simulator, CLASS. It takes ~ 30 seconds on average to do one forward simulation. For example, in our application, it took 520 minutes to generate the targets $(D,\,q,\,P_{\text{lin}})$ for 1000 input cosmologies. For each of the output, we assume the following model

$$\boldsymbol{y}=\boldsymbol{\Phi\beta}+\boldsymbol{f}+\boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\Phi}$ consist of a set of basis functions, $\boldsymbol{\beta}$ is a set of latent variables for which we assume a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\mathbf{C}$, $\boldsymbol{f}$ is the deterministic error component of the model and $\boldsymbol{\epsilon}$ is just the noise term, with zero mean and covariance $\boldsymbol{\Sigma}$. A multivariate

Gaussian prior is used for $\boldsymbol{f}$, with zero mean and covariance (kernel matrix) $\mathbf{K}$. In particular, for the latter, we choose the Squared-Exponential kernel function such that

$$\text{cov}(f_i, f_j) = \lambda^2 \text{exp}[-\frac{1}{2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^{\text{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)]. \tag{4}$$

with λ and $\boldsymbol{\Omega} = \text{diag}(\omega_1, \omega_2 \ldots \omega_5)$ being the set of kernel hyperparameters. These are fixed by maximising the marginal likelihood:

$$\log p(\boldsymbol{y}) = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu})^{\text{T}}(\mathbf{K}_y + \boldsymbol{\Phi}\mathbf{C}\boldsymbol{\Phi}^{\text{T}})^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}) - \frac{1}{2}\log\left|\mathbf{K}_y + \boldsymbol{\Phi}\mathbf{C}\boldsymbol{\Phi}^{\text{T}}\right| + \text{constant} \tag{5}$$

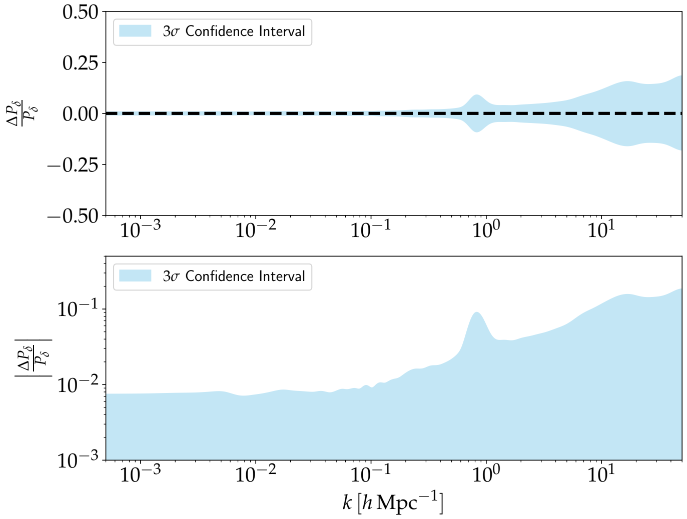


Figure 1: To investigate the performance of the emulator, we draw an independent set of cosmological parameters, randomly from the prior and we calculate the fractional error between the predicted ones with the surrogate model and CLASS. The mean of ${}^{\Delta P_\delta}/_{P_\delta}$ is shown by the broken horizontal line and the 3σ confidence interval, derived from the standard deviations of all experiments, is shown in pale blue. For an accurate emulator, it is expected that the mean is centred on 0 and this demonstrates the robustness of this method. Note that in this procedure, one can also specify the number of desired power spectra for $z \in [0.0, 5.0]$. For example, for p cosmological parameters and n redshifts, we have np power spectra outputs. In the bottom panel, we show the absolute error on a logarithmic scale.

where we have defined $\mathbf{K}_y = \mathbf{K} + \boldsymbol{\Sigma}$. For this application, we use polynomial basis functions, $[1,\, \boldsymbol{\theta},\, \boldsymbol{\theta}^2]$ and a similar approach was adopted by Schneider et al. (2011) who used a second order polynomial model for emulating the CBM power spectrum. Training the emulator, that is, learning the kernel hyperparameters, for the different targets, took around 340 minutes. All experiments were conducted on an Intel Core i7-9700 CPU desktop computer. Once the models are trained and stored, the mean $\bar{y}_* = \mathbf{X}_*\hat{\boldsymbol{\beta}} + f_*$ and variance $\text{var}(y_*) = \mathbf{X}_*\mathbf{V}_\beta\mathbf{X}_*^{\text{T}} + k_{**} + \sigma_*^2 - \boldsymbol{k}_*^{\text{T}}\mathbf{K}_y^{-1}\boldsymbol{k}_*$ can be computed given a test point, $\boldsymbol{\theta}_*$ within the prior range. We have defined $\mathbf{X}_* = \boldsymbol{\Phi}_* - \boldsymbol{k}_*^{\text{T}}\mathbf{K}_y^{-1}\boldsymbol{\Phi}$ and $f_* = \boldsymbol{k}_*^{\text{T}}\mathbf{K}_y^{-1}\boldsymbol{y}$. $\boldsymbol{\Phi}_*$ is the set of basis functions computed at the test point, $\boldsymbol{\theta}_*$. $\hat{\boldsymbol{\beta}} = \mathbf{V}_{\boldsymbol{\beta}}[\boldsymbol{\Phi}^{\text{T}}\mathbf{K}_y^{-1}\boldsymbol{y} + \mathbf{C}^{-1}\boldsymbol{\mu}]$ and $\mathbf{V}_\beta = [\boldsymbol{\Phi}^{\text{T}}\mathbf{K}_y^{-1}\boldsymbol{\Phi} + \mathbf{C}^{-1}]^{-1}$ correspond to the posterior mean and variance of $\boldsymbol{\beta}$. In addition, the first and second derivatives of the surrogate model are:

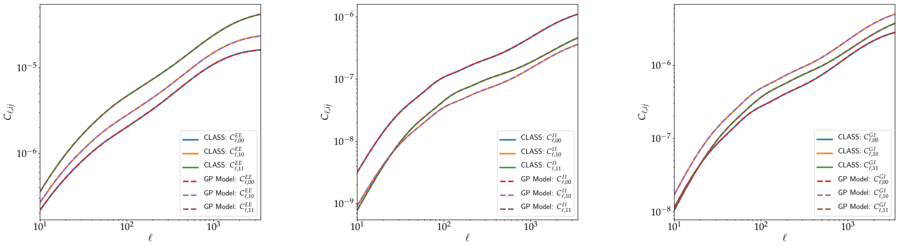
$$\frac{\partial \bar{y}_*}{\partial \boldsymbol{\theta}_*} = \frac{\partial \boldsymbol{\Phi}_*}{\partial \boldsymbol{\theta}_*}\hat{\boldsymbol{\beta}} + \left[\boldsymbol{k}_* \odot \mathbf{Z}_*\boldsymbol{\Omega}^{-1}\right]^{\text{T}}\mathbf{K}_y^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{\beta}}) \tag{6}$$

and

$$\frac{\partial^2 \bar{y}_*}{\partial \boldsymbol{\theta}_*^2} = \frac{\partial^2 \boldsymbol{\Phi}_*}{\partial \boldsymbol{\theta}_*^2} \hat{\boldsymbol{\beta}} + \left[\Omega^{-1} \frac{\partial \boldsymbol{k}_*}{\partial \boldsymbol{\theta}_*} \mathbf{Z}_* - \Omega^{-1} \odot \boldsymbol{k}_* \right] \mathbf{K}_y^{-1} (\boldsymbol{y} - \boldsymbol{\Phi} \hat{\boldsymbol{\beta}}). \tag{7}$$

3 Results

In this section, we highlight the main results after the different approximate models are constructed. In particular, we use 1000 training points to build the latter and we use a separate, independent set of 100 power spectra computed using the simulator to assess the performance of the emulator.



(a) $C^{\mathrm{EE}}_{\ell,ij}$

(b) $C^{\mathrm{II}}_{\ell,ij}$

(c) $C^{\mathrm{GI}}_{\ell,ij}$

Figure 2: The left, centre and right panels show the different weak lensing power spectra as calculated by the emulator (broken curves) and the accurate model, CLASS, shown by the solid curves. The different power spectra within each panel correspond to the auto- and cross- power spectra, due to the 2 tomographic redshift distributions, hence leading to 00, 10, and 11 power spectra. These power spectra are then added, via the intrinsic alignment parameter, A_{IA} to construct a final model, $C^{\mathrm{tot}}_{\ell,ij}$ in a weak lensing analysis.

In Figure 1, the top and bottom panel show the fractional uncertainty and absolute fractional uncertainty in linear and logarithmic scales respectively. In Figure 2, we show the different types of weak lensing power spectra calculated using CLASS and the emulator. The left, middle and right panel show the auto- and cross- EE, II and II power spectra due to the two Gaussian tomographic bins. In the three panels, the blue, orange and green curves correspond to the auto- and cross- power spectra, $C_{\ell,00}$, $C_{\ell,10}$ and $C_{\ell,11}$ as computed by CLASS. Similarly, the red, purple and brown broken curves are the power spectra generated by the emulator. The power spectra are in agreement when comparing CLASS and the emulator. Note that, in a typical weak lensing analysis, the three different types of power spectra (EE, GI and II) are combined together via the intrinsic alignment parameter, A_{IA}, that is, $C^{\mathrm{tot}}_{\ell,ij} = C^{\mathrm{EE}}_{\ell,ij} + A^2_{\mathrm{IA}} C^{\mathrm{II}}_{\ell,ij} - A_{\mathrm{IA}} C^{\mathrm{GI}}_{\ell,ij}$.

We also test the emulator on simulated weak-lensing bandpowers. We assume measurements over $10 \leq \ell \leq 1500$ and 5 tomographic slices with Gaussian $n(z)$, centred on redshifts [0.5, 1.0, 1.5, 2.0, 2.5] and each having a standard deviation of 0.075. Ten bandpowers, equally spaced in logarithmic scale, are used and this gives us a set of 150 data points. Moreover, we simulate and then assume in the likelihood independent Gaussian errors with, for simplicity, $\sigma = 0.5\hat{\mathcal{B}}_\ell$, where $\hat{\mathcal{B}}_\ell$ is the bandpower evaluated at the fiducial set of cosmological parameters. For this particular case, we have set $A_{\mathrm{IA}} = 0$ but one can trivially include this factor and marginalise over it in the sampling process. The fiducial point $\boldsymbol{\theta}_{\mathrm{fid}} = [0.12, 0.0225, 3.45, 1.0, 0.72]$ is used to generate the data and is shown by the black dots in Figure 3. We use a Gaussian likelihood and uniform priors on all cosmological parameters, similar to the range of the inputs of the emulator. Figure 3 shows the results obtained when sampling the cosmological parameters on this toy data set. The red contours correspond to the result using the emulator while the pale blue colour refers to the posterior distributions using CLASS. We run three separate MCMC chains using the `emcee` sampler (Foreman-Mackey et al., 2013), each with 150 000 MCMC samples, two with the emulator and one with CLASS. On each of the three resulting pairs of runs, we compute the Gelman-Rubin convergence parameter, $\hat{R}$ (Gelman & Rubin, 1992). The worst $\hat{R}$ value is 1.002, consistent with all three chains being drawn from the same distribution, and corroborating the agreement shown in Figure 3. The emulator developed in this work is thus able to robustly recover the posterior distributions of all the cosmological parameters, compared to the accurate solver, CLASS.

4 Conclusions

In this work, we have introduced a new method for accelerating the computation of the 3D matter power spectrum. The emulator is around 300 times faster compared to the full simulator and this can further be increased if leveraging better hardware. We have shown that the different weak lensing power spectra can also be computed and we have showcase one application of the emulator to a Bayesian inverse problem in cosmology. This work can be used further in other applications, for example, in the case where we want to use the Hamiltonian Monte Carlo (HMC) sampler, which requires derivatives. The latter is also important in the calculation of the Fisher information matrix. Moreover, methods such as approximate inference based on Taylor series expansion also demand for derivatives, which can be computed in a fast way, as shown in this work. A challenging problem might be in scaling this method but techniques such as Robust Bayesian Committee Machine (rBCM) can be explored (Deisenroth & Ng, 2015). The code and documentation for this work are available at https://github.com/Harry45/emuPK and https://emupk.readthedocs.io/ respectively.

Broader Impact

Current weak lensing surveys only cover part of the sky and with the development of data-intensive surveys such as the KiDS (de Jong et al., 2013), Vera C. Rubin Observatory (Almoubayyed et al., 2020), Euclid (Laureijs et al., 2011), HSC (Aihara et al., 2018) and DES Abbott et al. (2016) which will cover a large part of the sky, the tool developed in this work can easily be integrated in future weak lensing data analysis pipelines. Moreover, N-body simulation codes are generally very expensive and the method behind the approximate model can be used to model the Universe.

References

Abbott, T., Abdalla, F. B., Allam, S., Amara, A., & Dark Energy Survey Collaboration. 2016, Phys. Rev. D, 94, 022001

Agarwal, S., Abdalla, F. B., Feldman, H. A., Lahav, O., & Thomas, S. A. 2012, MNRAS, 424, 1409

Agarwal, S., Abdalla, F. B., Feldman, H. A., Lahav, O., & Thomas, S. A. 2014, MNRAS, 439, 2102

Aihara, H., Arimoto, N., Armstrong, et al. 2018, PASJ, 70, S4

Almoubayyed, H., Mandelbaum, R., Awan, H., et al. 2020, MNRAS, 499, 1140

Alsing, J., Charnock, T., Feeney, S., & Wandelt, B. 2019, MNRAS, 488, 4440

Alsing, J. & Wandelt, B. 2019, MNRAS, 488, 5093

Alsing, J., Wandelt, B., & Feeney, S. 2018, MNRAS, 477, 2874

Aricò, G., Angulo, R. E., & Zennaro, M. 2021, arXiv e-prints, arXiv:2104.14568

Auld, T., Bridges, M., Hobson, M. P., & Gull, S. F. 2007, MNRAS, 376, L11

de Jong, J. T. A., Verdoes Kleijn, G. A., Kuijken, K. H., & Valentijn, E. A. 2013, Experimental Astronomy, 35, 25

Deisenroth, M. P. & Ng, J. W. 2015, arXiv e-prints, arXiv:1502.02843

Fendt, W. A. & Wandelt, B. D. 2007, ApJ, 654, 2

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Gelman, A. & Rubin, D. B. 1992, Statistical Science, 7, 457

Habib, S., Heitmann, K., Higdon, D., Nakhleh, C., & Williams, B. 2007, Phys. Rev. D, 76, 083503

Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156

Heitmann, K., Lawrence, E., Kwan, J., Habib, S., & Higdon, D. 2014, ApJ, 780, 111

Heitmann, K., White, M., Wagner, C., Habib, S., & Higdon, D. 2010, ApJ, 715, 104

Ho, M.-F., Bird, S., & Shelton, C. R. 2021, arXiv e-prints, arXiv:2105.01081

Köhlinger, F., Viola, M., Joachimi, B., et al. 2017, MNRAS, 471, 4412

Laureijs, R., Amiaux, J., Arduini, S., Auguères, J. L., et al. 2011, arXiv e-prints, arXiv:1110.3193

Lawrence, E., Heitmann, K., White, M., et al. 2010, ApJ, 713, 1322

Lesgourgues, J. 2011, arXiv e-prints, arXiv:1104.2932

Lewis, A. & Challinor, A. 2011, CAMB: Code for Anisotropies in the Microwave Background

Schneider, M. D., Holm, Ó., & Knox, L. 2011, ApJ, 728, 137

Appendix

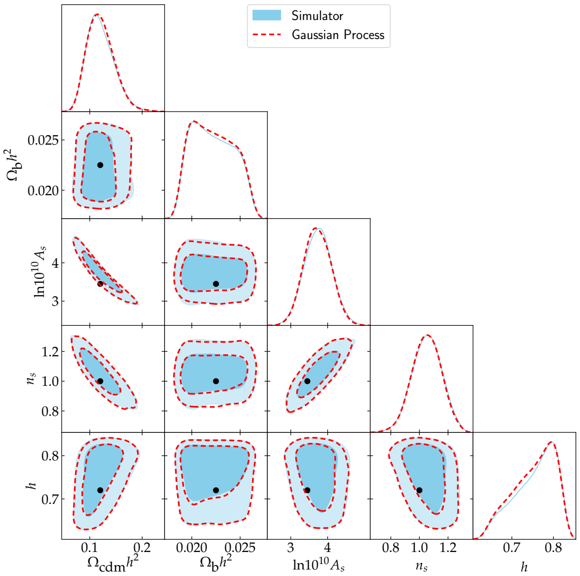


Figure 3: Marginalised posterior distributions of the five cosmological parameters. The blue color refers to the posterior distribution of the parameters as inferred using CLASS and the broken red contours refer to the posterior distribution when using the emulator developed in this work. The black dots correspond to the fiducial point in parameter space where the data have been generated.